NAVAL POSTGRADUATE SCHOOL

MONTEREY, CALIFORNIA

THESIS

OPTIMIZING CAT-ASVAB ITEM SELECTION USING FORM ASSEMBLY TECHNIQUES

by

Toby Lee

June 2006

Thesis Advisor:	Robert F. Dell
Second Reader:	Johannes O. Royset

Approved for public release; distribution is unlimited

THIS PAGE INTENTIONALLY LEFT BLANK

REPORT DOCUMENTATION PAGE			Form Approved OMB No. 0704-0188
Public reporting burden for this collection of information is estimated to average 1 hour per response, including the time for reviewing instruction, searching existing data sources, gathering and maintaining the data needed, and completing and reviewing the collection of information. Send comments regarding this burden estimate or any other aspect of this collection of information, including suggestions for reducing this burden, to Washington headquarters Services, Directorate for Information Operations and Reports, 1215 Jefferson Davis Highway, Suite 1204, Arlington, VA 22202-4302, and to the Office of Management and Budget, Paperwork Reduction Project (0704-0188) Washington DC 20503.			
1. AGENCY USE ONLY ***(Leave blank)***	**2. REPORT DATE** June 2006	**3. REPORT TYPE AND DATES COVERED** Master's Thesis	
4. TITLE AND SUBTITLE: Optimizing CAT-ASVAB Item Selection Using Form Assembly Techniques			**5. FUNDING NUMBERS**
6. AUTHOR(S) Lee, Toby			
7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES) Naval Postgraduate School Monterey, CA 93943-5000			**8. PERFORMING ORGANIZATION REPORT NUMBER**
9. SPONSORING /MONITORING AGENCY NAME(S) AND ADDRESS(ES) DMDC 400 Gigling Rd. Seaside, CA 93955			**10. SPONSORING/MONITORING AGENCY REPORT NUMBER**
11. SUPPLEMENTARY NOTES The views expressed in this thesis are those of the author and do not reflect the official policy or position of the Department of Defense or the U.S. Government.			
12a. DISTRIBUTION / AVAILABILITY STATEMENT Approved for public release; distribution is unlimited			**12b. DISTRIBUTION CODE**

13. ABSTRACT (maximum 200 words)

The Armed Services Vocational Aptitude Battery (ASVAB) is a test that approximately 700,000 students in 12,000 high schools take each year to determine military occupation placement. Form Assembly for the ASVAB refers to the selection of 20-35 questions, known as items, from an item pool of approximately 300 items to create a paper and pencil test in one of its ten topics. Previous research formulates form assembly as an Integer Linear Program (ILP). The current ASVAB mostly uses a Computer Adaptive Test (CAT), which estimates an examinee's ability after the examinee answers each item and selects the next item based on prior performance. The current CAT-ASVAB implementation does not control the number of items selected from each subject (taxonomy group) for a test. This thesis introduces ILPs, previously used for form assembly, that impose taxonomy restrictions and applies them to the CAT-ASVAB. We create four ILP variations and test them against the current method of item selection, by simulating 3,500 examinees (500 examinees each for seven given ability levels). The results show that all of the ILPs have acceptable solution times for CAT use, and taxonomy restrictions can be imposed while also having more even exposure rates (the number of times an item is administered divided by the number of examinees) than the current implementation of the CAT-ASVAB. A variation that relaxes most of the binary variables and constrains the difficulty of each item to be within a predetermined magnitude of the current ability estimate, performs the best in terms of item exposure (for both under and over-utilized items) and error between an examinee's estimated ability level and actual ability level.

14. SUBJECT TERMS Computer Adaptive Test, ASVAB, Form Assembly, Integer Linear Programming, Optimization, Operations Research			**15. NUMBER OF PAGES** 57
			16. PRICE CODE
17. SECURITY CLASSIFICATION OF REPORT Unclassified	**18. SECURITY CLASSIFICATION OF THIS PAGE** Unclassified	**19. SECURITY CLASSIFICATION OF ABSTRACT** Unclassified	**20. LIMITATION OF ABSTRACT** UL

NSN 7540-01-280-5500 Standard Form 298 (Rev. 2-89)
Prescribed by ANSI Std. 239-18

THIS PAGE INTENTIONALLY LEFT BLANK

Approved for public release; distribution is unlimited

OPTIMIZING CAT-ASVAB ITEM SELECTION USING FORM ASSEMBLY TECHNIQUES

Toby T. Lee
Civilian, Defense Manpower Data Center
B.A., Unversity of California, Berkeley, 1999

Submitted in partial fulfillment of the
requirements for the degree of

MASTER OF SCIENCE IN OPERATIONS RESEARCH

from the

NAVAL POSTGRADUATE SCHOOL
June 2006

Author: Toby Lee

Approved by: Robert F. Dell
Thesis Advisor

Johannes O. Royset
Second Reader

James N. Eagle
Chairman, Department of Operations Research

THIS PAGE INTENTIONALLY LEFT BLANK

ABSTRACT

The Armed Services Vocational Aptitude Battery (ASVAB) is a test that approximately 700,000 students in 12,000 high schools take each year to determine military occupation placement. Form Assembly for the ASVAB refers to the selection of 20-35 questions, known as items, from an item pool of approximately 300 items to create a paper and pencil test in one of its ten topics. Previous research formulates form assembly as an Integer Linear Program (ILP). The current ASVAB mostly uses a Computer Adaptive Test (CAT), which estimates an examinee's ability after the examinee answers each item and selects the next item based on prior performance. The current CAT-ASVAB implementation does not control the number of items selected from each subject (taxonomy group) for a test. This thesis introduces ILPs, previously used for form assembly, that impose taxonomy restrictions and applies them to the CAT-ASVAB. We create four ILP variations and test them against the current method of item selection, by simulating 3,500 examinees (500 examinees each for seven given ability levels). The results show that all of the ILPs have acceptable solution times for CAT use, and taxonomy restrictions can be imposed while also having more even exposure rates (the number of times an item is administered divided by the number of examinees) than the current implementation of the CAT-ASVAB. A variation that relaxes most of the binary variables and constrains the difficulty of each item to be within a predetermined magnitude of the current ability estimate, performs the best in terms of item exposure (for both under and over-utilized items) and error between an examinee's estimated ability level and actual ability level.

THIS PAGE INTENTIONALLY LEFT BLANK

TABLE OF CONTENTS

THIS PAGE INTENTIONALLY LEFT BLANK

LIST OF FIGURES

THIS PAGE INTENTIONALLY LEFT BLANK

LIST OF TABLES

THIS PAGE INTENTIONALLY LEFT BLANK

ACKNOWLEDGMENTS

I would like to express my deepest gratitude to my advisor, Professor Robert Dell for his guidance, time, and support throughout my research. His efforts were invaluable in bringing this thesis to a successful completion.

I would also like to thank Professor Johannes Royset, my second advisor, for taking the time to provide his insight and support for me to complete this thesis.

I am also grateful to Iosif Krass and Mary Pommerich at the Defense Manpower Data Center for offering me valuable resources necessary for me to complete my research, and to the Defense Manpower Data Center for giving me the opportunity to work on this topic.

Finally, I would like to thank my family and friends for their continued encouragement. Without their support, I almost certainly would not have come this far.

THIS PAGE INTENTIONALLY LEFT BLANK

EXECUTIVE SUMMARY

The Armed Services Vocational Aptitude Battery (ASVAB) is a test that approximately 700,000 students in 12,000 high schools take each year to determine military occupation placement. Form Assembly for the ASVAB refers to the selection of 20-35 questions, known as items, from an item pool of approximately 300 items to create a paper and pencil test in one of its ten topics. ASVAB form assembly has been previously formulated as an integer linear program (ILP) with an objective function that minimizes the deviation from a predetermined goal curve for the test.

Most of the ASVAB tests are administered as a Computer Adaptive Test (CAT). The CAT estimates an examinee's ability after the examinee answers each item and selects the next item based on prior performance. Because the CAT is able to determine an examinee's ability level after each question and select future questions based on this estimator, the test length for a CAT is shorter than a paper and pencil test. However, the current CAT-ASVAB does not control the number of items selected from each subject (taxonomy group) for a test. Therefore, this taxonomy distribution of the items in a test can be heavily skewed toward a particular subject. A solution to this problem is for a test to not only select the next item, but select an entire test trajectory for the examinee's current estimated ability. This is called a shadow test, and this thesis combines a shadow test with previously researched paper and pencil form assembly for application to the CAT-ASVAB.

This thesis also discusses other problems associated with the CAT, such as item exposure control and solution time. One method it explores is item-stratification. In this method, the item selection algorithm divides the item pool into groups according to their discrimination parameter (an item with a high discrimination parameter is able to separate examinees with nearly the same ability, whereas a low discrimination parameter does not separate them as well) and divides the test into an equal number of stages. The purpose is

to select items with a lower discrimination (and therefore lower information value) toward the beginning of a test, and leave items with a higher discrimination (and higher information value) until the end when the ability estimate is more accurate.

There are five variations of CAT-ASVAB item selection considered in this thesis: 1) A previously researched paper and pencil form assembly method for the ASVAB (KM); 2) KM that constrains the difficulty parameter (a parameter that measures the difficulty of an item) to be within a certain amount of the current ability level of the examinee (DM); 3) KM with the addition of item-stratification constraints (SM); and 4) KM that has both difficulty parameter constraints and item stratification constraints (SDM); 5) The current item selection method of the CAT-ASVAB (OM), is a benchmark to compare the other four. Each of the five variations of the model is examined using 3,500 artificially generated examinees (500 examinees each for seven given ability levels). Aside from SM and SDM having a high maximum exposure rate, our results indicate that all of the shadow test variations have more even exposure rates than the current implementation of the CAT-ASVAB, having significantly less unutilized items. DM performs the best in terms of item exposure (for both under and over-utilized items) and error between an examinee's estimated ability level and actual ability level. All of the variations benefit from the ability to add taxonomy constraints. Without the taxonomy constraints, our results suggest that the current CAT implementation has a taxonomy distribution heavily favoring one of the taxonomy groups.

I. INTRODUCTION

Since 1968, all US military applicants take the Armed Services Vocational Aptitude Battery (ASVAB) to determine military occupation placement. Approximately 700,000 students in 12,000 High Schools take this test every year [Pommerich 2005]. Form assembly for the ASVAB refers to the selection of multiple choice questions, known as items, out of a given item pool to create a paper and pencil test in one of its ten topics. A typical form has 20-35 items selected from an item pool of approximately 300 items. Kunde [1997] formulates form assembly as an integer linear program (ILP) and solves it both optimally and using heuristics.

In 1997, many ASVAB tests were still commonly administered in their printed (paper and pencil) form. The ASVAB has since moved toward being a Computer Adaptive Test (CAT) [e.g., Weiss 2004]. Other tests that use a CAT include the GRE [e.g., Syvum 2006] and GMAT [e.g., Princeton Review 2006]. The CAT estimates an examinee's ability after the examinee answers each item and selects the next item based on this estimator. This allows it to use fewer items than a paper and pencil exam to determine an examinee's ability.

The current CAT-ASVAB item selection algorithm does not currently take into account item taxonomy constraints [Sands, Waters, and McBride 1999]. A taxonomy constraint imposes a limit on the number of items from a given subject (e.g. Addition, Division, etc.). Veldkamp and van der Linden (2004) use a shadow test to determine the next question. A shadow test creates a whole test trajectory for the examinee's current estimated ability then chooses the best item amongst that trajectory to administer. By creating this whole test, other constraints can be added to the formulation, including taxonomy constraints.

This thesis extends the ILP from Kunde [1997] for use as a shadow test and applies it to item selection for a CAT-ASVAB. The primary extensions speed solution time and control item exposure. Item exposure control refers to limiting the number of

times a test administers an item to a set of examinees. Too many examinees receiving the same item increases the likelihood of a future examinee having advanced knowledge of an item.

II. BACKGROUND

A. TEST THEORY

The ASVAB uses Item Response Theory (IRT) to measure the precision of each test. An examinee's ability level is denoted as θ. It is assumed that θ follows a standard normal distribution (mean of zero and a standard deviation of one). The range of θ is commonly set between -3.0 and 3.0 or -2.5 or 2.5 [Sands, Waters, and McBride 1999]. In IRT, the probability of an examinee, with ability level θ, answering an item correctly is calculated with the three parameter logistic function shown below [Lord 1980]:

$$p(\theta) = c + \frac{1-c}{1+e^{-Da(\theta-b)}}.$$

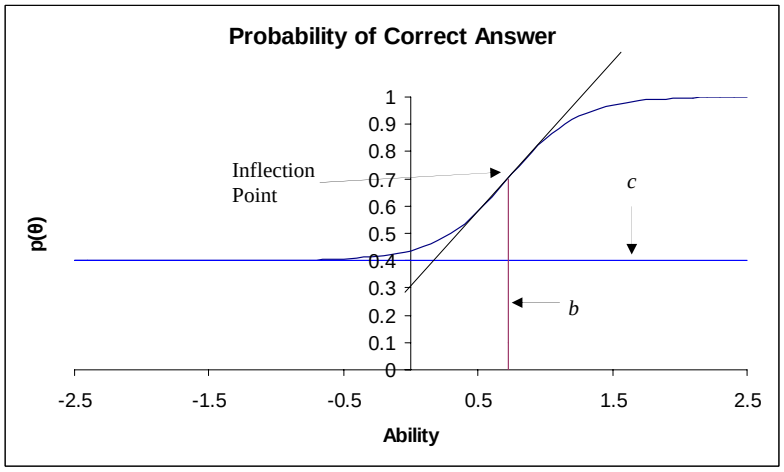

Figure 1: Sample Logistic Function
In the above sample, the discrimination parameter: a=2.24, the difficulty parameter: b=0.72, and the guessing parameter: c=0.4

The 3 parameters are *a*, *b*, and *c*, with *D* being a scaling factor. The *a* parameter is the discrimination of the item. This is the capability of the item to distinguish between applicants of different abilities. In Figure 1, the *a* parameter is proportional to the slope of the logistic function at its inflection point. The steeper the slope, the greater the difference examinees with different ability levels have in answering an item correctly; a flatter slope means examinees with different ability levels have more similar probabilities of a correct response. The *b* parameter measures the difficulty of an item. In Figure 1,

the *b* parameter determines the position of the curve's inflection point along the θ-axis. Finally, parameter *c* is the guessing parameter. This is the probability of a person with a low ability level guessing the item correctly. This parameter shows up in Figure 1 as the lower asymptotic bound on *p(θ)*'s axis. These parameters are typically calculated after the item has been pretested 1,000 to 10,000 times. From here, the item information function can be derived from *p(θ)*, [Lord 1980]

$$I_i(\theta) = \frac{p'(\theta)}{p(\theta)(1-p(\theta))}, \text{ or}$$

$$I_i(\theta) = \frac{D^2a^2(1-c)}{(c+e^{Da(\theta-b)})(1+e^{-Da(\theta-b)})},$$

where p' is the derivative of p. The presence of the derivative in the numerator indicates that items with a higher discrimination parameter have a higher information value. Because the information contribution of each item is assumed to be independent of the other items in the ASVAB, the item information functions can be added together to produce an overall information curve. With *N* being the number of items in the form, the exam information function is [Lord 1980]:

$$I(\theta) = \sum_{i=1}^{N} I_i(\theta).$$

This function measures the precision of the exam in estimating an examinee's true ability level. The next section shows how the above information function is applicable to form assembly.

B. OPTIMIZATION OF FORM ASSEMBLY FOR ASVAB (PAPER AND PENCIL)

This section describes Kunde's paper and pencil formulation, which is used in the optimization model in this thesis for the CAT-ASVAB. Kunde's formulation has two goals expressed in the objective function. The first is to minimize the difference between

the information of the exam and the information from a goal curve. A goal curve is a test information function like the one introduced in the previous section that represents the desired information distribution of the exam across the ability levels. It is produced from empirical research and testing. The deviations between an assembled form and the goal curve for specific values of θ are organized by their magnitude into groups which are denoted in the formulation below by the index *g*. Each group is assigned a penalty per unit of deviation. Higher deviations from the goal curve receive a higher penalty per unit deviation.

For security purposes, alternate forms are created for an exam (denoted by the index *f*). This leads to the second goal of the formulation: to make each form as similar as possible in information. The second component of the objective function seeks to minimize the deviations of each form from the first reference form.

Below is Kunde's integer linear program formulation for the paper and pencil form.

Indices:

i	item from the item pool;
θ	ability level;
f	form to be assembled (1,2,…F);
t	taxonomy(1,2,…T);
g	penalty group

Sets:

$TaxItems_t$	The set of items in taxonomy group t

Data:

CAT_g	The maximum deviation between a form and the goal curve in group g
$INF_{i\theta}$	Information value of item i at percentile θ
$NITEM_t$	The required number of items in taxonomy t
$PARAWEI$	Weight that combines the two goals

$PENALTY_g$	Penalty per unit deviation within group g
$SHAPE_\theta$	The information value for the goal curve at percentile θ

Variables:

x_{if}	One, if item i is used in form f
$py_{\theta fg}$	Deviation above the goal curve in group g at percentile θ on form f
$ny_{\theta fg}$	Deviation below the goal curve in group g at percentile θ on form f
$delplus_f$	The total information form one contains that exceeds form f
$delneg_f$	The total information form f contains that exceeds form one

Formulation:

min

$$\sum_\theta \sum_f \sum_g PENALTY_g (py_{\theta fg} + ny_{\theta fg}) + PARAWEI \sum_{f>1} (delplus_f - delneg_f) \qquad \text{(k1)}$$

such that

$$\sum_g py_{\theta fg} \geq \sum_i INF_{i\theta} x_{if} - SHAPE_\theta \qquad \forall \theta, f \qquad \text{(k2)}$$

$$\sum_g ny_{\theta fg} \geq -\sum_i INF_{i\theta} x_{if} + SHAPE_\theta \qquad \forall \theta, f \qquad \text{(k3)}$$

$$\sum_{i \in TaxItem_t} x_{if} = NITEM_t \qquad \forall f, t \qquad \text{(k4)}$$

$$\sum_f x_{if} \leq 1 \qquad \forall i \qquad \text{(k5)}$$

$$\sum_i \sum_\theta INF_{i\theta} x_{i1} - \sum_i \sum_\theta INF_{i\theta} x_{if} = delplus_f - delneg_f \qquad \forall f > 1 \qquad \text{(k6)}$$

$$0 \leq py_{\theta fg} \leq CAT_g \qquad \forall \theta, f, g \qquad \text{(k7)}$$

$$0 \leq ny_{\theta fg} \leq CAT_g \qquad \forall \theta, f, g \qquad \text{(k8)}$$

x_{if} binary $\qquad \forall i, f \qquad$ (k9)

$delplus_f$, $delneg_f \geq 0 \qquad \forall f \qquad$ (k10)

The first component in the objective function (k1), corresponding to the first goal of minimizing deviation from the goal curve, $\sum_{\theta}\sum_{f}\sum_{g} PENALTY_g(py_{\theta fg} + ny_{\theta fg})$, expresses the vertical deviation from the goal curve. The variables $py_{\theta fg}$ and $ny_{\theta fg}$ are the positive and negative deviations, respectively, of form *f* from the goal curve, in group *g*, for ability θ. In the second component of the objective function, $PARAWEI\sum_{f>1}(\ delplus_f + delneg_f)$, the variable $delplus_f$ is the total form one information in excess of form *f*, while $delneg_f$ is the total form *f* information in excess of form one.

Constraints sets (k2) and (k3) give the values for the positive and negative deviations of the information function from the goal curve. Set (k4) specifies the number of items in a form from a given taxonomy. Set (k5) states that item *i* can only appear in at most one form. Set (k6) gives the total difference in information between the forms, and sets (k7) and (k8) bound the deviations of the information function from the goal curve.

C. CAT-ASVAB

The formulation above optimizes the objective function across all θs, and creates a form that satisfies a set of specified attributes (e.g., length and taxonomy). In a CAT, the examinee's current performance on the exam determines each item that is administered. Therefore, at a given point in an exam, an individual with a higher ability level receives an item of more difficulty than an individual with a lower estimated ability. Because the examinee receives an item based on his estimated ability, the exam can produce a better estimate for the examinee's ability in fewer questions. As currently implemented, all examinees start with the same average ability level estimate, $\theta_0 = 0$. The CAT-ASVAB uses the Owens Bayes algorithm of calculating the ability after each item is answered. Because the order of items administered affects the ability calculation, an additional Bayesian module calculation is used to calculate θ at the end of the test. Currently, the item selection algorithm for the CAT-ASVAB seeks to maximize the item

information function at the examinee's current θ and limit item exposure. The information values are pulled from a table by θ. [Sands, Waters, and McBride 1999]

1. Shadow Test

One method proposed to deal with the taxonomy constraints is a shadow test [e.g. van der Linden and Veldkamp 1998]. Instead of merely calculating the best item to administer at the current θ, a whole test trajectory is constructed for the examinee at the current θ. The indices used in the formulation below are the same as in Kunde's formulation with the addition of an index *h*, a quantitative attribute group. An example of a quantitative attribute group is the total word count for all items in the group adding up to a pre-specified total. Thus, a possible constraint would be to limit the total word count for a set of items in each group *h*. This is represented by the following constraints:

$$\sum_{i \in Q_h} L_{ih} x_i \leq UH_h \qquad \forall h$$

$$\sum_{i \in Q_h} L_{ih} x_i \geq LH_h \qquad \forall h,$$

where L_{ih}, in this example, is the word count for item *i*, UH_h and LH_h are an upper and lower bound respectively on the sum of the word counts for all items in group *h*, and Q_h is the set of items in group *h*. Below is Veldkamp and van der Linden's formulation using notation consistent with Kunde's formulation above.

Indices:

k	iteration count where examinee is given his *k*th question
h	quantitative attribute group

Sets:

Fix	The set of items already administered
Q_h	The set of items in quantitative attribute group *h*

Data:

$\hat{\theta}_{k-1}$ Current ability estimate after *k*-1 items have been administered

L_{ih} Quantitative attribute for item *i* for attribute group *h*

UH_h Upper bound for number of items in group *h*

LH_h Lower bound for number of items in group *h*

UT_t Upper bound for number of items in taxonomy *t*

LT_t Lower bound for number of items in taxonomy *t*

$I_i(\theta)$ The item information value at θ

Decision Variable:

x_i One, if item *i* is used in the shadow test

Formulation:

$$\max \sum_i I_i(\hat{\theta}_{k-1}) x_i \qquad \text{(v1)}$$

such that

$$x_i = 1 \qquad \forall i \in Fix \qquad \text{(v2)}$$

$$\sum_{i \in TaxItems_t} x_i \le UT_t \qquad \forall t \qquad \text{(v3)}$$

$$\sum_{i \in TaxItems_t} x_i \ge LT_t \qquad \forall t \qquad \text{(v4)}$$

$$\sum_{i \in Q_h} L_{ih} x_i \le UH_h \qquad \forall h \qquad \text{(v5)}$$

$$\sum_{i \in Q_h} L_{ih} x_i \ge LH_h \qquad \forall h \qquad \text{(v6)}$$

$$\sum_i x_i = N \qquad \text{(v7)}$$

x_i binary $\qquad \forall i \qquad$ (v8)

The model selects the item with the greatest information from the items in the shadow test that have not already been administered at the current ability, $\hat{\theta}_{k-1}$. Constraint set (v2) sets x_i to 1 for the items *i* that have already been administered. Constraint sets (v3) and (v4) are taxonomy constraints and set an upper and lower limit

respectively on the number of items administered from each taxonomy group. Constraint sets (v5) and (v6) are the above mentioned quantitative attribute constraints. "Because each shadow test meets the constraints, the adaptive test automatically meets them" [van der Linden and Veldkamp 2004].

2. Taxonomy and Item Exposure Control Research for CAT

Much research has been done on different ways to implement CAT. Because one of the main concerns with CAT is item exposure control, many papers written about CAT implementation discuss possible solutions for this issue. The CAT-ASVAB currently uses Sympson and Hetter's [1985] algorithm to control item exposure. This thesis uses this algorithm for its optimization model as well. The Sympson and Hetter algorithm assigns a number between zero and one, called the item exposure parameter, to each item. A pretest simulation determines these parameters. Items with a higher exposure rate at the end of the simulation receive a lower exposure parameter. During the actual test, when the test selects an item, it generates a random number uniformly distributed between zero and one. If the item exposure parameter of this item is less than the random number, the test rejects the item and selects the item with the next highest information value, and so on.

Another technique to control item exposure is called 5-4-3-2-1 [Sympson and Hetter 1997]. The first item is chosen randomly out of the five most informative items. The next item is then chosen randomly out of the four most informative, and so on until it is choosing from one item. Afterwards, the procedure starts over again at five items. Another randomization technique is to choose one item out of three, then disqualify the other two from further administration [Thomasson 1998]. Another technique does not use the item information value, but randomly selects from items within a specified distance from a target difficulty level [Lunz and Stahl 1998].

Other methods require a more significant change in item or test structure to address item exposure control. One method is item stratification, and this thesis also includes this method into its optimization model. Items fall into *n* groups called strata by

their *a* parameters, and exams divide into *n* stages. For a model with taxonomy constraints, this first categorizes the items by their taxonomy before sorting the items within each taxonomy by the *a* parameter. It then divides the items in each taxonomy into *n* groups. Items from the first group in each taxonomy go into the first strata, items from the second group go into the second strata, and so on until there are *n* strata. During the *n*th stage, the test selects an item from the *n*th strata [Leung, Chang, and Hau 2003]. Item stratification selects items with a lower discrimination value near the beginning of the test. Because items with a higher discrimination also carry higher information values, item stratification is contrary to the typical approach of selecting the item with the highest information value. Item stratification reserves the items that carry more information toward the end of the exam where the ability estimate is closer to the true ability. In a study done by Chang and van der Linden, item stratification yields more even exposure rates throughout the items, thus having fewer underexposed and overexposed items. Below is the formulation of the item stratification model into a shadow test. The indices are the same as the shadow test formulation given in the previous section, with the addition of the index *r*, the stratum. [Chang and van der Linden, 2003]

Indices:

r	stratum;

Sets:

Q_{r_k}	The set of items at the strata *r* when selecting item *k*

Data:

S_r	The required number of items from strata *r*
B_i	Difficulty of item *i* (standard deviations from θ=0)

Variables:

y	Deviation of item's difficulty parameter from $\hat{\theta}_{k-1}$

Formulation:

$\min y$ (c1)

such that

$(B_i - \hat{\theta}_{k-1})\, x_i \le y$ $\forall i \in Q_{r_k}$ (c2)

$(B_i - \hat{\theta}_{k-1})\, x_i \ge -y$ $\forall i \in Q_{r_k}$ (c3)

$x_i = 1$ $\forall i \in Fix$ (c4)

$$\sum_{i \in Q_{r_k}} x_i = S_r \qquad \forall r \qquad \text{(c5)}$$

$$\sum_{i \in TaxItems_t} x_i \le UT_t \qquad \forall t \qquad \text{(c6)}$$

$$\sum_{i \in TaxItems_t} x_i \ge LT_t \qquad \forall t \qquad \text{(c7)}$$

$$\sum_{i \in Q_h} L_{ih} x_i \le UH_h \qquad \forall h \qquad \text{(c8)}$$

$$\sum_{i \in Q_h} L_{ih} x_i \ge LH_h \qquad \forall h \qquad \text{(c9)}$$

$y \ge 0$ (c10)

x_i binary $\forall i$ (c11)

Items with a difficulty parameter closest to the current estimate of ability, $\hat{\theta}_{k-1}$, are chosen within the given constraints. Constraint set (c4) specifies the number of items that must come from each strata. The rest of the constraints are the same as the shadow test.

Another method, the Computerized Adaptive Sequential Test (CAST), partitions the test into a collection of subtests such that these subtests become the units of test administration instead of items [Davis and Dodd 2003]. This method groups the items into subtests called modules and places them in multistaged panels. There are two ways to construct the panels. The first is bottom-up construction that assembles the items into modules such that each module, as a self-contained unit, meets the requisite information,

content, and item feature targets selected for the test [Davis and Dodd 2003]. The second method of panel construction is top-down, where any module path through the panel results in a test of appropriate precision, content, and item type [Davis and Dodd 2003]. The method used in Davis and Dodd's study is the bottom-up construction. With the exception of the first stage, the test segregates the modules by difficulty level in each stage. The first stage has only one module. A typical allocation for the other stages would place three modules in the second and third stage, with each module corresponding to a low, medium, and high difficulty. A panel is randomly assigned to an examinee at the beginning. From there, at the first stage, the examinee receives a subtest. When the examinee completes the module, the test calculates his ability, and in the next stage, it bases the next module the examinee receives on his current estimated ability. An examinee can only move up one level between stages. For example, one cannot receive an easy module after completing a hard module the stage before. Like *a*-stratification, this method also yielded more even exposure rates [David and Dodd 2003].

Two of the methods mentioned above for item exposure control, the Sympson and Hetter algorithm and item stratification, are incorporated into the optimization model for this thesis as well as alternate forms from the paper and pencil exam. Shadow tests in the existing research use the existing maximum information or minimum difficulty deviation as objective functions. The formulation in the following section, however, uses the deviation from a goal curve as in Kunde's paper and pencil formulation for the objective function.

THIS PAGE INTENTIONALLY LEFT BLANK

III. THE CAT-ASVAB OPTIMIZATION MODELS

A. SHADOW TEST FORMULATION AND VARIATIONS

The integer linear program (ILP) in this thesis uses Kunde's formulation as a starting point and adapts it for use in the CAT-ASVAB as a shadow test. In his paper and pencil formulation, Kunde uses alternative forms as a means of test security. This shadow test formulation retains the alternative forms as a means of item exposure control along with the Sympson-Hetter method. For this thesis, the test creates two forms, with 15 items each, for each shadow test. An examinee starts off on one of the forms. Each item selected first goes through the Sympson-Hetter algorithm. If the algorithm rejects an item, the test administers the item with the most information from the alternative form. The test does not use the rejected items again for the remainder of the exam. If the Sympson-Hetter algorithm also rejects the item from the alternative form, the test goes back and selects the next most informative item from the first form, and so on. If the items in the shadow tests to choose from run out, the test reruns the model to obtain a new shadow test.

As mentioned earlier, the solution time of the shadow test is critical. To speed up solution times, this formulation relaxes Kunde's ILP such that only the x_{if} value for the current item needs to be binary, while the rest of the x_{if} values can be continuous. Allowing continuous variables could decrease overall solution quality, but we did not observe any substantial differences. For the relaxation, the formulation splits x_{if} into a binary and continuous component, xb_{if} and xc_{if}, respectively. Therefore the constraint set from the original formulation:

$$x_{if} \text{ binary} \qquad \forall i, f$$

is replaced with the below constraint sets.

$$0 \le x_{if} \le 1 \qquad \forall i, f$$

$$0 \le xc_{if} \le 1 \qquad \forall i, f$$

$$xb_{if} \text{ binary} \qquad \forall i, f$$

$$x_{if} = xb_{if} + xc_{if} \qquad \forall i, f$$

To specify that at least one x_{if}, other than the administered items, is an integer, the following constraint is added.

$$\sum_i xb_{if} \ge \sum_{i \in Fix} x_{if} + 1 \qquad \forall f$$

Kunde's formulation, along with the addition of the above constraints, establishes the base model for this thesis (KM). For this thesis, we develop three other variations for comparison. One variation (DM) comes from the observation that items administered with a higher deviation between the *b* parameter and current ability estimate tend to have a smaller effect on the ability estimate. For example, if an individual answered an item correctly in which the difficulty parameter was far below his current ability, it would barely affect the new ability estimate. Therefore, for this variation, the two constraints below are added to constrain the difficulty parameter to be within a given number, *BLIM*, of the current ability estimate.

$$(b_i - \hat{\theta}_{k-1})\, x_{if} \le BLIM \qquad \forall i \notin Fix$$

$$(b_i - \hat{\theta}_{k-1})\, x_{if} \ge -BLIM \qquad \forall i \notin Fix$$

Using the same notation as Kunde's formulation and van der Linden's sample shadow test, below is the formulation for this variation.

Data:

BLIM Maximum deviation of item difficulty from current ability

Variables:

x_{if}	One, if item *i* is used in form *f*
xc_{if}	Continuous component of x_{if}
xb_{if}	Binary component of x_{if}

Formulation:

min

$$\sum_{\theta}\sum_{f}\sum_{g} PENALTY_g(py_{\theta fg} + ny_{\theta fg}) + PARAWEI\sum_{f>1}(delplus_f - delneg_f) \qquad \text{(d1)}$$

s.t.

$$\sum_{g} py_{\theta fg} \geq \sum_{i} INF_{i\theta} x_{if} - SHAPE_{\theta} \qquad \forall \theta, f \qquad \text{(d2)}$$

$$\sum_{g} ny_{\theta fg} \geq -\sum_{i} INF_{i\theta} x_{if} + SHAPE_{\theta} \qquad \forall \theta, f \qquad \text{(d3)}$$

$$\sum_{i \in TaxItem_t} x_{if} = NITEM_t \qquad \forall f, t \qquad \text{(d4)}$$

$$\sum_{f} x_{if} \leq 1 \qquad \forall i \qquad \text{(d5)}$$

$$(b_i - \hat{\theta}_{k-1})\, x_{if} \leq BLIM \qquad \forall i \notin Fix, f \qquad \text{(d6)}$$

$$(b_i - \hat{\theta}_{k-1})\, x_{if} \geq -BLIM \qquad \forall i \notin Fix, f \qquad \text{(d7)}$$

$$\sum_{i}\sum_{\theta} INF_{i\theta} x_{i1} - \sum_{i}\sum_{\theta} INF_{i\theta} x_{if} = delplus_f - delneg_f \qquad \forall f > 1 \qquad \text{(d8)}$$

$$0 \leq py_{\theta fg} \leq CAT_g \qquad \forall \theta, f, g \qquad \text{(d9)}$$

$$0 \leq ny_{\theta fg} \leq CAT_g \qquad \forall \theta, f, g \qquad \text{(d10)}$$

$$x_{if} = 1 \qquad \forall i \in Fix, f \qquad \text{(d11)}$$

$$x_{if} = xb_{if} + xc_{if} \qquad \forall i, f \qquad \text{(d12)}$$

$$\sum_{i} xb_{if} \geq \sum_{i \in Fix} x_{if} + 1 \qquad \forall f \qquad \text{(d13)}$$

$$0 \leq x_{if} \leq 1 \qquad \forall i, f \qquad \text{(d13)}$$

$$0 \leq xc_{if} \leq 1 \qquad \forall i, f \qquad \text{(d15)}$$

xb_{if} binary $\forall i, f$ (d16)

$delplus_f$, $delneg_f \geq 0$ $\forall f$ (d17)

The second variation (SM) uses item stratification. It adds the below constraint, adapted from Chang and van der Linden's shadow test formulation with item stratification, to the formulation.

$$\sum_{i \in Q_r} x_{if} = S_r \qquad \forall r, f$$

In order to ensure that the decision variable for an item from the current stage is binary, the formulation sets all of the items in the shadow test at the current stage as binary. The below constraint achieves this purpose.

$$\sum_{i \in Q_r} xb_{if} = S_r \qquad \forall r = CURSTG, f$$

where *CURSTG* is the current stage of the exam.

The third variation (SDM) combines the DM and SM formulations. However, instead of adding the two constraints to limit the difficulty parameter, the formulation relaxes the two constraints and inserts them into the objective function as a price for deviating too far from the current ability estimate. The new objective function is therefore

$$\min \sum_{\theta} \sum_{f} \sum_{g} PENALTY_g (py_{\theta fg} + ny_{\theta fg}) + PARAWEI \sum_{f>1} (delplus_f - delneg_f)$$

$$+ DIFPEN \sum_{i} \sum_{f} \left(pbdev_{if} + nbdev_{if} \right)$$

where $pbdev_{if}$ and $nbdev_{if}$ are given below

$$(b_i - \hat{\theta}_{k-1})\, x_{if} \leq BLIM + pbdev_{if} \qquad \forall i \notin Fix, f$$

$$(b_i - \hat{\theta}_{k-1})\, x_{if} \geq - BLIM - nbdev_{if} \qquad \forall i \notin Fix,\ f\ ,$$

and *DIFPEN* is the penalty per unit for more than *BLIM* units over or under the current ability estimate. The reason for not adding the difficulty constraints directly into the formulation is because combined with the item stratification constraints, the addition of the difficulty parameter constraints tends to result in an infeasible solution. Below is the SDM formulation.

Data:

CURSTG Current stage of exam

Variables:

$pbdev_{if}$ The additional positive deviation of item *i*'s difficulty parameter from the current ability estimate greater than *BLIM*

$nbdev_{if}$ The additional negative deviation of item *i*'s difficulty parameter from the current ability estimate less than *BLIM*

Formulation:

Min

$$\sum_{\theta}\sum_{f}\sum_{g} PENALTY_g (py_{\theta fg} + ny_{\theta fg}) + PARAWEI \sum_{f>1} (delplus_f - delneg_f)$$

$$+ DIFPEN \sum_{i}\sum_{f} \left(pbdev_{if} + nbdev_{if} \right) \qquad \text{(sd1)}$$

s.t.

$$\sum_{g} py_{\theta fg} \geq \sum_{i} INF_{i\theta} x_{if} - SHAPE_{\theta} \qquad \forall \theta, f \qquad \text{(sd2)}$$

$$\sum_{g} ny_{\theta fg} \geq -\sum_{i} INF_{i\theta} x_{if} + SHAPE_{\theta} \qquad \forall \theta, f \qquad \text{(sd3)}$$

$$\sum_{i \in TaxItem_t} x_{if} = NITEM_t \qquad \forall f, t \qquad \text{(sd4)}$$

$$\sum_{f} x_{if} \leq 1 \qquad \forall i \qquad \text{(sd5)}$$

$$(b_i - \hat{\theta}_{k-1})\, x_{if} \leq BLIM + pbdev_{if} \qquad \forall i \notin Fix, f \qquad \text{(sd6)}$$

$$(b_i - \hat{\theta}_{k-1})\, x_{if} \geq - BLIM - nbdev_{if} \qquad \forall i \notin Fix, f \qquad \text{(sd7)}$$

$$\sum_{i} \sum_{\theta} INF_{i\theta} x_{i1} - \sum_{i} \sum_{\theta} INF_{i\theta} x_{if} = delplus_f - delneg_f \qquad \forall f > 1 \qquad \text{(sd8)}$$

$$0 \leq py_{\theta fg} \leq CAT_g \qquad \forall \theta, f, g \qquad \text{(sd9)}$$

$$0 \leq ny_{\theta fg} \leq CAT_g \qquad \forall \theta, f, g \qquad \text{(sd10)}$$

$$x_{if} = 1 \qquad \forall i \in Fix, f \qquad \text{(sd11)}$$

$$x_{if} = xb_{if} + xc_{if} \qquad \forall i, f \qquad \text{(sd12)}$$

$$\sum_{i \in Q_r} xb_{if} = S_r \qquad \forall r = CURSTG, f \qquad \text{(sd13)}$$

$$\sum_{i \in Q_r} x_{if} = S_r \qquad \forall r, f \qquad \text{(sd14)}$$

$$0 \leq x_{if} \leq 1 \qquad \forall i, f \qquad \text{(sd15)}$$

$$0 \leq xc_{if} \leq 1 \qquad \forall i, f \qquad \text{(sd16)}$$

$$xb_{if} \text{ binary} \qquad \forall i, f \qquad \text{(sd17)}$$

$$delplus_f,\ delneg_f \geq 0 \qquad \forall f \qquad \text{(sd18)}$$

B. ABILITY CALCULATION

The Owens Bayes algorithm [Sands, Waters, and McBride 1999], which the CAT-ASVAB normally uses to calculate the ability after an examinee answers each item, assumes that if an examinee answers an item correctly, he receives a more difficult item next, and if he answers incorrectly, he receives an easier item [Krass 2005]. Because none of the shadow test variations above consistently follow this behavior, this thesis uses a different algorithm to estimate the ability after an examinee answers each item.

This algorithm, developed by Dan Segall of DMDC, unlike the Owens Bayes algorithm, is independent of the order the test administers the items and whether or not the test administers an item of higher difficulty to an examinee after a correct answer [Krass 2005]. Calculation time is slower than the Owens Bayes algorithm, but it is still within 30 seconds, which is our criterion for an acceptable solution time for a CAT [Krass 2005].

THIS PAGE INTENTIONALLY LEFT BLANK

IV. RESULTS OF CAT-ASVAB OPTIMIZATION SIMULATIONS

A. SETUP FOR SIMULATION

To test the performance of the model, we run simulations for each shadow test variation. GAMS [GAMS 2006] generates all integer linear programs (ILP) and XA [Sunset 2003] solves them on a 1.7 GHz Dell workstation. We use a similar approach to Chang and van der Linden's paper on item stratification and select a few ability levels for the simulations. Those ability levels are θ=−1.5, −1.0, −0.5, 0, 0.5, 1.0, and 1.5. For each of these ability levels, the simulation creates 500 examinees. Each examinee takes a test generated by each of the five variations. The first is the current implementation of the CAT-ASVAB, which administers items by maximum information (OM). This is the benchmark for comparing the other four variations. The other four shadow test variations come from a CAT-ASVAB optimization formulation: the variation derived from Kunde's paper and pencil formulation adapted for the CAT (KM), the variation with constraints on the difficulty parameters (DM), the variation using item stratification (SM), and the variation with item stratification and difficulty parameter constraints (SDM).

Discretization of ability levels provide information only for those values of θ selected. But we have high confidence for those ability levels. This discretization also corresponds to an underlying assumption that examinee ability levels follow a uniform distribution. An alternative strategy would be to sample from a continuous distribution (for example, the standard normal). Previous CAT research has observed that sampling from a continuous distribution of θ would imply using enormous sample sizes to get reasonable estimates of the bias and mean squared error (MSE) functions, which still would have to be pooled over classes of θ values and be accurate only near the center of the distribution [Chang and van der Linden, 2003]. There are two consequences from this assumption. "First, the results for the bias and MSE functions are conditional on θ [Chang and van der Linden, 2003]." But, because the accuracy of these functions are not dependent on the distribution of the examinees, one can generalize the results for the bias and MSE to any population of examinees. "Second, the results for the item exposure rates do not necessarily generalize to other populations of examinees [Chang and van der Linden, 2003]."

The item pool contains approximately 170 items and comes from the Mathematical Knowledge test for the CAT-ASVAB [Sands, Waters, and McBride 1999]. These items are an experimental set and are not an actual item pool currently in use for the CAT-ASVAB. Each shadow test variation has about 2,500 constraints, 350 binary variables, and 2,000 continuous variables.

The initial ability estimate for each test variation is $\theta = 0$. After the simulated examinees take the tests, the simulation outputs a set of deviations between the true and estimated θ for each examinee. Then using S-Plus 6.2 [Insighful 2003], we run a Wilcoxon Sign-Rank Test to compare each shadow test variation's deviation distribution to OM [e.g. Conover 1999]. Table 1 gives the parameters used for the shadow tests.

For all Shadow Test Variations	
Forms per Shadow Test	2
Number of Items per Form	15
Scaling Factor (*D*)	1.7
Number of items required from taxonomy group 1 ($NITEM_1$)	2
Number of items required from taxonomy group 2 ($NITEM_2$)	4
Number of items required from taxonomy group 3 ($NITEM_3$)	8
Number of items required from taxonomy group 4 ($NITEM_4$)	1
For DM and SDM	
Maximum allowable deviation of difficulty from current ability (*bLimit)*	0.5
For SM and SDM	
Number of items required from strata 1 (S_1)	3
Number of items required from strata 2 (S_2)	4
Number of items required from strata 3 (S_3)	4
Number of items required from strata 4 (S_4)	4
Repetitions (or number of examinees)	500

Table 1: Parameter Settings for Formulations
There are five variations altogether (the four shadow test variations and OM) with 3,500 repetitions for each (500 repetition for seven given ability levels).

B. RESULTS

Table 2 shows the taxonomy distribution for the simulations. The simulation altogether selects 52,500 items (15 items for each of the 3,500 tests) for each test variation. OM performs poorly in terms of the taxonomy constraints specified. A majority of items administered in the OM simulation come from taxonomy group 3. This

is most likely because in the item pool, 103 of the 170 items are in taxonomy group 3. On the other hand, the four shadow test variations follow the taxonomy constraints shown in Table 1.

Taxonomy		
Group	KM, DM, SM, and SDM	OM
1	7000	2858
2	14000	6028
3	28000	40150
4	3500	3464

Table 2: Taxonomy Distribution
Taxonomy distribution for OM heavily favors taxonomy group 3, while the taxonomy distribution for KM, DM, SM, and SDM follow the parameters set by the simulation (shown in Table 1)

Table 3 shows the solution times of each shadow test variation. The times include the program generation, runtime, and output time for GAMS. KM and DM have acceptable results with maximum solution times under 10 seconds. The item stratification variations, SM and SDM, however, have higher maximum solution time. The long solution time occurred primarily at the selection of the 12^{th} item, which is the beginning of the 4^{th} and final stage. With the exception of that item, solution times are as quick as the other variations for the selection of the rest of the items in the test. If needed, the maximum solution times could possibly be reduced by using direct problem generation or another solver. But, we do not explore these options in this thesis.

Solution Time (seconds)			
Shadow Test Variation	Max	Min	Average
KM	7.731	0.24	0.472
DM	3.245	0.27	0.522
SM	1036.66	0.27	1.865
SDM	189.012	0.34	3.924

Table 3: Solution Times
The solution time for KM, DM, SM, and SDM, on average, is acceptable. But, the high maximum solution times for SM and SDM make them infeasible options.

Figure 2 shows the exposure rates of the items for each variation. They are calculated by dividing the number of times the item is administered by the number of tests. The *x*-axis lists the items in descending order according to their exposure rates.

Although SM and SDM start off much higher, all of the shadow test variations eventually end up approaching a more uniform distribution than OM. OM has the highest amount of unused items at 77 items. SDM and SM have the next highest number of unused items at 37 and 34 items respectively. Of even more concern, however, are the extremely high exposure rates with SM carrying a maximum exposure rate of 1 and SDM carrying a maximum exposure rate of 0.86. The problem items, although different for each variation, are distributed at the start of the exam. A possible reason for this is that items at the beginning of the test have a lower discrimination. So their Sympson and Hetter parameters are very high (close to or equal to 1), making the test much less likely to reject the items. Therefore, the Sympson and Hetter algorithm would rarely reject an item at the first stage. KM and DM administered all of the items in their simulations. As the graph shows, the curves for KM and DM have the flattest slopes, which indicate low maximum exposure rates and low number of unutilized items.

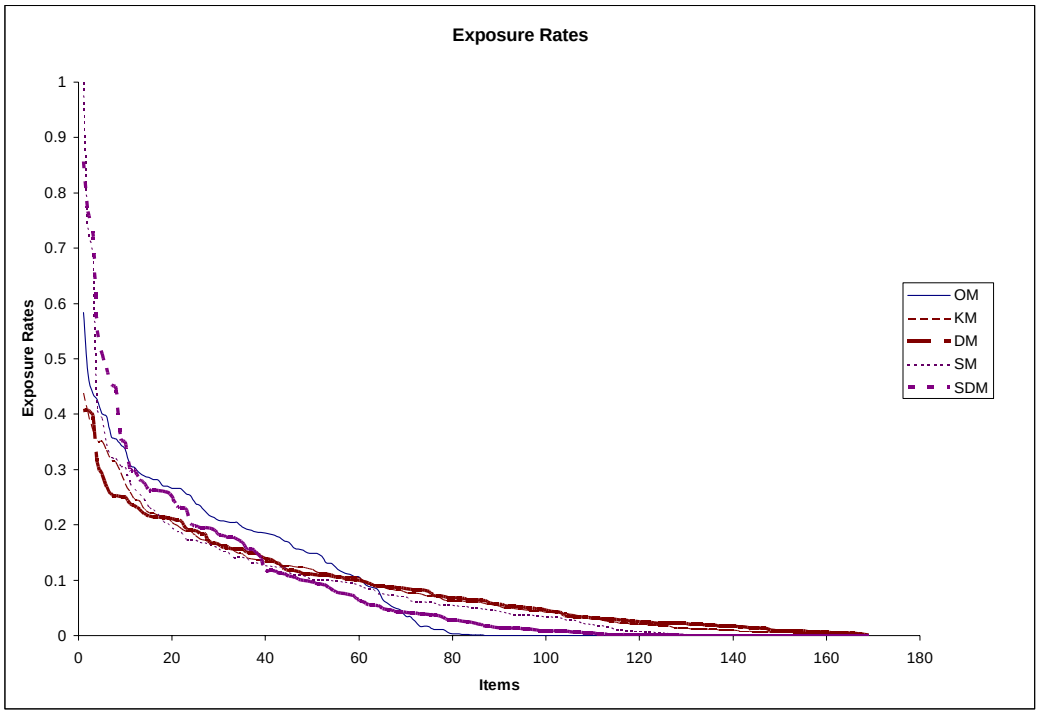

Figure 2: Exposure Rates:
OM is given by a solid line. KM is given by a thin dashed line. DM is given by a bold dashed line, SM is given by a thin dotted line, and SDM is given by a bold dotted line. The *x*-axis lists the items in descending order according to their exposure rates.

Figures 3-7 below are the histograms of the errors for each test variation. The error for each examinee's estimated ability is:

$$\hat{\theta}_k - \theta_k ,$$

where $\hat{\theta}_k$ is the estimated ability level of examinee *k* after the exam, and θ_k is examinee *k*'s true ability level. There are 3,500 examinees for each test variation (500 examinees for each of the seven pre-selected ability levels). The Wilcoxon Sign Rank test p-values are given in Table 4. For this simulation, we use a two-sided test to determine whether there is a difference between the mean and medians of each shadow test variation's deviation distribution to that of OM. Using a 90% Confidence Interval, a p-value of under 0.05 would indicate a significant difference between the means and medians of a given formulation against OM. The p-values for DM and SDM are equal to zero. Therefore, DM and SDM differ significantly from OM.

p-values overall			
KM	DM	SM	SDM
0.1417	0	0.2489	0

Table 4: p-values versus OM for Wilcoxon Sign Rank Test
DM and SDM significantly differ from OM because their p-values are below 0.05.

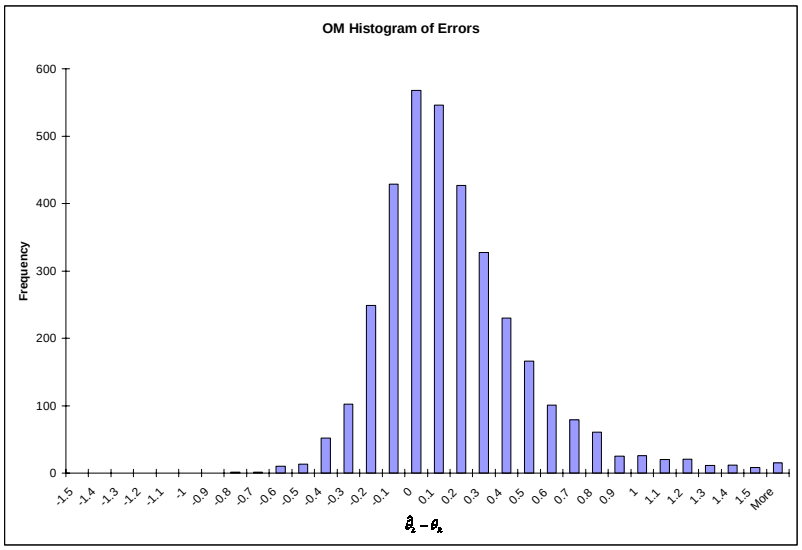

Figure 3: Error Histogram of OM
The *x*-axis gives the error range for θ (given by $\hat{\theta}_k - \theta_k$); The *y*-axis gives the frequency for the errors

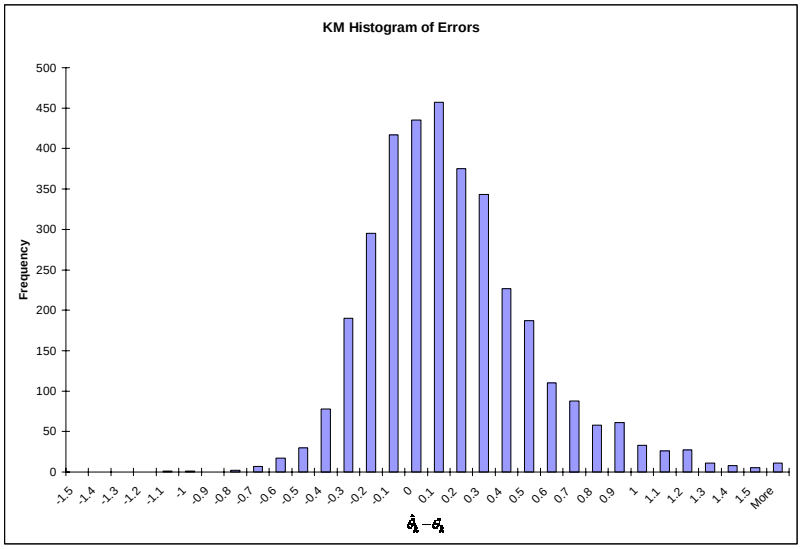

Figure 4: Error Histogram of KM

The *x*-axis gives the error range for θ (given by $\hat{\theta}_k - \theta_k$); The *y*-axis gives the frequency for the errors

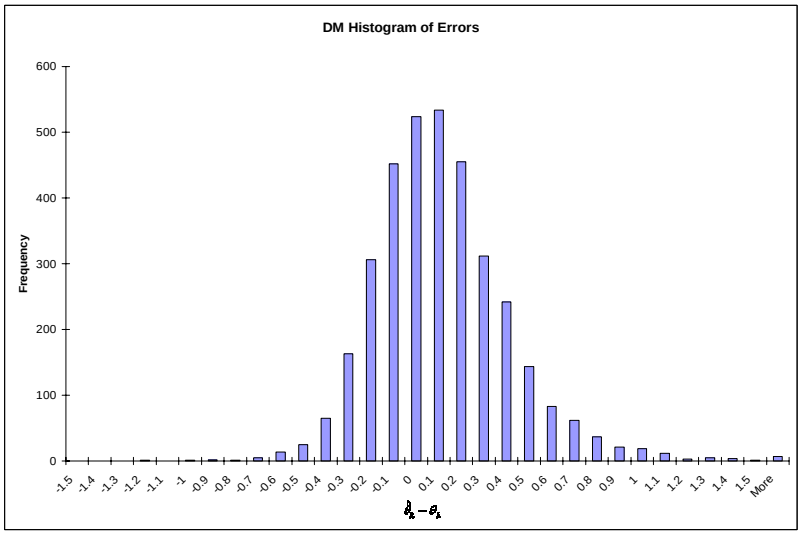

Figure 5: Error Histogram of DM

The *x*-axis gives the error range for θ (given by $\hat{\theta}_k - \theta_k$); The *y*-axis gives the frequency for the errors

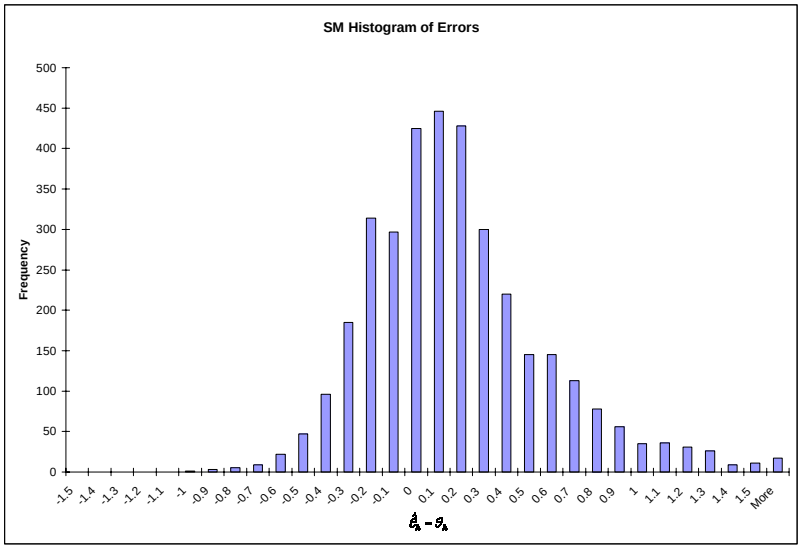

Figure 6: Error Histogram of SM

The *x*-axis gives the error range for θ (given by $\hat{\theta}_k - \theta_k$); The *y*-axis gives the frequency for the errors

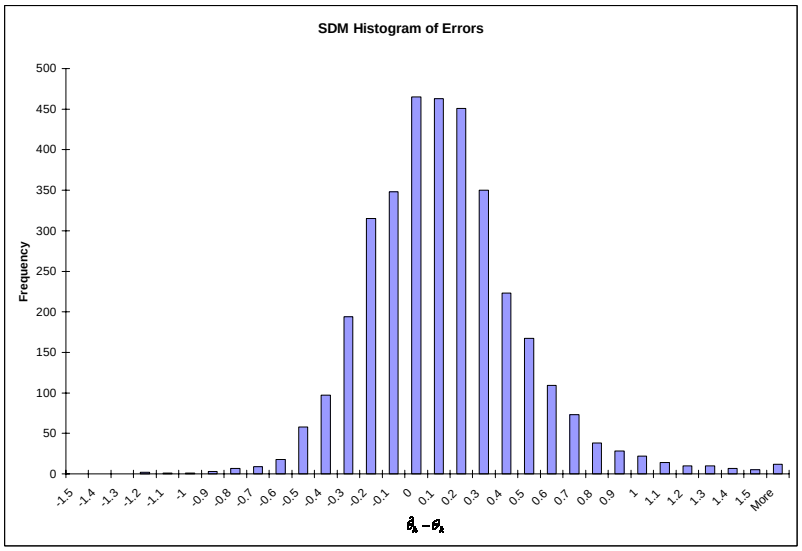

Figure 7: Error Histogram of SDM

The *x*-axis gives the error range for θ (given by $\hat{\theta}_k - \theta_k$); The *y*-axis gives the frequency for the errors

Figures 8 and 9 below show the bias and mean squared error (MSE) functions. The values in the graphs are discrete with polynomial interpolation (from MS Excel) to obtain the intermediate values. In terms of the bias functions, each test variation

performs similarly with a large bias for more extreme negative ability levels. The graphs are also consistent with the results from the Wilcoxon Sign Rank Test. The KM and SM curves have steep slopes like the OM curve at the extreme negative values of θ. OM performs better than KM and SM for most of the curve, and performs better than all of the shadow test variations at $\theta \geq 0.5$. This is not surprising as there are no taxonomy constraints on OM. The two variations that were shown to be significantly different than OM, DM and SDM, have a flatter slope and do not have the steep negative slope at the extreme negative ability levels. Of particular note, DM performs better than OM for most of the curve at $\theta < 0.5$. Also, with the exception of $\theta = -0.5$ where the magnitude of the bias is only slightly higher than that of OM, SDM performs better than OM at the same regions as DM.

Because the bias functions for each variation behave similarly, it is not surprising that the MSE functions for each variation do as well, with large errors as θ approaches the extreme negative values. OM performs the best for most of the curve, $\theta \geq 0$, and performs better than KM and SM for all values of θ. Like the bias curve, the MSE curves for DM and SDM are flatter than OM, and therefore perform better at extreme negative values of θ, with DM's MSE lower than SDM's MSE for the whole curve.

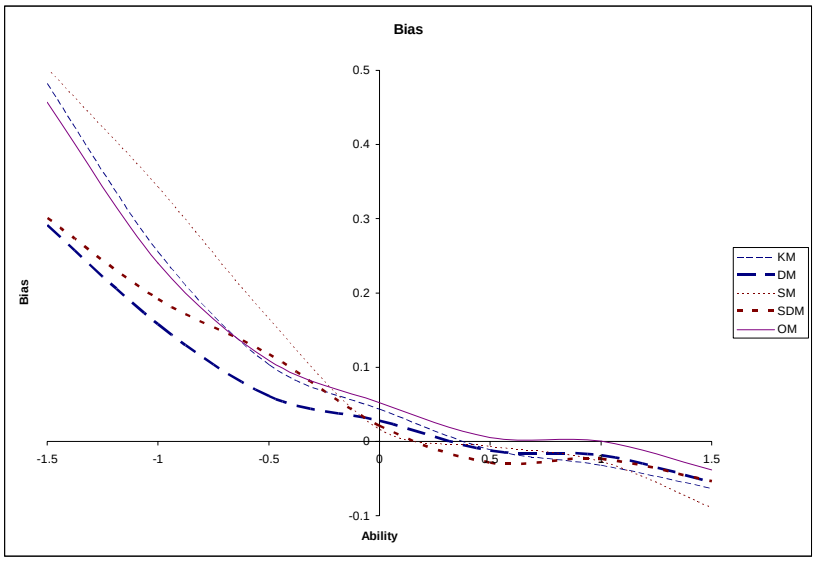

Figure 8: Bias Function:
OM is given by a solid line. KM is given by a thin dashed line. DM is given by a bold dashed line, SM is given by a thin dotted line, and SDM is given by a bold dotted line.

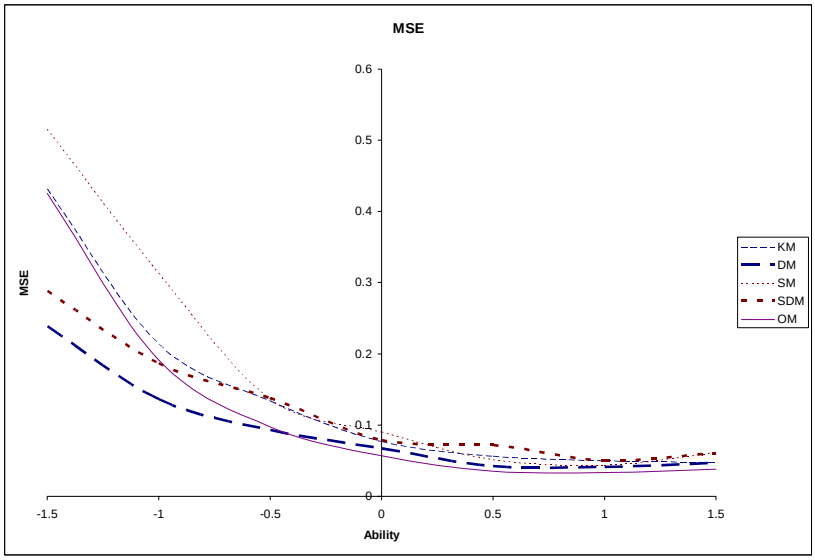

Figure 9: MSE Function:
OM is given by a solid line. KM is given by a thin dashed line. DM is given by a bold dashed line, SM is given by a thin dotted line, and SDM is given by a bold dotted line.

THIS PAGE INTENTIONALLY LEFT BLANK

V. CONCLUSIONS AND FUTURE RESEARCH

A. CONCLUSIONS

The simulation results show that the current implementation of the CAT would benefit from the use of shadow tests. The primary motivation behind using the shadow tests for the CAT-ASVAB is to control taxonomy. This thesis introduces integer linear program (ILP) formulations that achieve this objective while our computational experience shows that the current method of item selection for the CAT-ASVAB (OM) has a taxonomy distribution that heavily favors one taxonomy group. In the area of item exposure, there are also significant benefits over OM. There are fewer unutilized items for each shadow test variation. In the case of the first and second shadow test variation (KM and DM), all items are administered, and maximum exposure rates are also lower than OM. The consequence of using the shadow test variations instead of OM is a slight loss in precision. As stated in Chang and van der Linden's paper, "the loss (in accuracy) can be made up for by adding a few items to the test, whereas the loss in credibility for a testing program due to item compromise or the financial loss involved in inefficient item usage is much more difficult to compensate [Chang and van der Linden, 2003]."

Given the five metrics for the simulation (bias, mean squared error (MSE), exposure rates, solution times, and taxonomy distribution), DM would be the most recommended amongst the shadow test variations. Like the rest of the shadow test variations, it meets the taxonomy constraints, with the solution time on average being the fastest. It actually has a lower bias for most of the curve than OM. Finally, the mean squared error (MSE) is the second lowest next to OM and even has a lower MSE at the negative values of θ. On the other hand, because of the high maximum exposure rates and maximum solution times, the shadow test variations with item stratification (SM and SDM) would not be recommended, despite also having a close bias and MSE to OM.

B. FUTURE RESEARCH

Because an experimental set of items comprises the item pool for this thesis simulation, further research can use an existing or future item pool to execute the

formulations. Also, only data for the Mathematical Knowledge (MK) test is used. Therefore item pools for the other CAT-ASVAB tests can be used in future research. Another area that can be extended is the sampling of the examinees. One could use a continuous distribution instead of sampling discrete values of θ. Also, this thesis only uses MSE and bias, whereas the current CAT-ASVAB uses the Birnbaum Score Information Function to measure precision of the exam [Sands, Waters, and McBride 1999]. Therefore, future research can also use this function.

LIST OF REFERENCES

Chang, H, van der Linden, W.J., 2003, *Implementing Content Constraints in Alpha-Stratified Adaptive Testing Using a Shadow Test Approach,* Applied Psychological Measurement, 27, 107-120.

Conover, W.J., 1999, *Practical Nonparametric Statistics,* John Wiley Inc., New York, NY.

Davis, L.L., Dodd, B.G., 2003, *Item Exposure Constraints for Testlets in the Verbal Reasoning Section of the MCAT,* Applied Psychological Measurement, 27, 335-356.

GAMS Development Corporation 2006, *GAMS – Release Notes*, http://www.gams.com/docs/release/release.htm, 4 January 2006, Washington, DC.

Insighful Corporation 2003, *S-Plus 6.2 for Windows Release Notes*, https://www.insighful.com, 19 May 2006, Seattle, Washington.

Krass, I., 2005, Defense Manpower Data Center, Mathematician, Personal correspondence, November 2005.

Kunde, D, 1997, *Optimization of Form Assembly for the Armed Services Vocational Aptitude Battery (ASVAB),* Masters Thesis in Operations Research, Naval Postgraduate School, Monterey, CA.

Leung, C., Chang H., Hau, K., 2003, *Incorporation of Content Balancing Requirements in Stratification Designs for Computerized Adaptive Testing.* Educational and Psychological Measurement, 63, 257-270.

Lord, F.M., 1980, *Applications of Item Response Theory to Practical Testing Problems,* Lawrence Erlbaum Associates, Hillsdale, NJ.

Lunz, M.E., Stahl, J.A., 1998, *Patterns of item exposure using a randomized CAT algorithm*, Paper presented at the annual meeting of the National Council of Measurement in Education, San Diego, CA.

Pommerich, M., 2005, Defense Manpower Data Center, Psychometrician, Personal correspondence, September 2005.

The Princeton Review, 2006, *GMAT: A Computer-Adaptive Test*, http://www.princetonreview.com, 19, May 2006, New York, NY.

Sands, W. A., Waters, B. K., McBride, J. R., 1999, *CATBOOK Computerized Adaptive Testing: From Inquiry to Operation*, Human Resources Research Organization, Alexandria, VA.

Sunset Software Technology, 2003, *XA Callable Library*, http://www.sunsetsoft.com/, 3 January 2006, San Marina, CA.

Sympson, J.B., Hetter, R.D., 1985, *Controlling Item Exposure Rates in Computerized Adaptive Tests*, Paper presented at the Annual Conference of the Military Testing Association, San Diego, CA.

Sympson, J.B., Hetter, R.D., 1997, *Item exposure control in CAT-ASVAB*. In Sands, W. A., Waters, B. K., McBride, J. R. (Eds.), Computerized Adaptive Testing: From Inquiry to Operation, 141-144.

Syvum Technologies Inc.,19 May 2006, *Computer-Adaptive Test for GRE: Information, http://www.syvum.com/gre/catgre.html*, London, ON, Canada.

Thommason, G.L., 1998, *CAT Item exposure control; New evaluation tools, alternate methods and integration into a total CAT program*, Paper presented at the annual meeting of the National Council of Measurement in Education, San Diego.

van der Linden, W. J., Veldkamp, B.P., 1998, *Optimal Test Assembly of psychological and Educational Tests*. Applied Psychological Measurement, 22, 195-211.

van der Linden, W.J., Chang, H., 2003, *Implementing Content Constraints in Alpha-Stratified Adaptive Testing Using a Shadow Test Approach*. Applied Psychological Measurement, 27, 107-120.

van der Linden, W. J., Veldkamp, B. P. 2004, *Constraining Item Exposure in Computerized Adaptive Testing with Shadow Tests*. Journal of Educational and Behavioral Statistics, 29, 273-291.

Weiss, D. J., 2004, *Computerized Adaptive Testing for Effective and Efficient Measurement in Counseling and Education*. Measurement and Evaluation in Counseling and Development, 37, 70-84.

Wilcoxon, F., 1945, *Individual Comparisons by Ranking Methods*. Biometrika, 1, 80-83.

THIS PAGE INTENTIONALLY LEFT BLANK

INITIAL DISTRIBUTION LIST

1. Defense Technical Information Center
   Ft. Belvoir, Virginia

2. Dudley Knox Library
   Naval Postgraduate School
   Monterey, California

3. Professor Robert F. Dell Code OR/DE
   Department of Operations Research
   Naval Postgraduate School
   Monterey, California

4. Professor Johannes O. Royset OR/DE
   Department of Operations Research
   Naval Postgraduate School
   Monterey, California